ISEBEL: Intelligent Search Engine for Belief Legends
HJ-253428-17

A collaboration among an international team of folklore scholars and computer scientists to develop analytical techniques for studying folkloric traditions across multiple national databases. The search engine can be reached at: https://search.isebel.eu

The performing groups were from the

- United States (PI: Timothy R. Tangherlini, Dept of Scandinavian, University of California, Los Angeles and Berkeley)
- The Netherlands (PI: Theo Meeder, The Meertens Institute of the Royal Netherlands Academy of Arts and Sciences)
- Germany (PIs: Christoph Schmitt, The Wossidlo Archive, Univ of Rostock; and Holger Meyer, Dept of Computer Science, Univ of Rostock).

Table of Contents:

1. Introduction

ISEBEL presents, for the first time ever, large scale data-driven computational research into traditional folk expressive culture across multiple tradition areas. To support this analytic work, we propose to develop an intelligent system that facilitates search, discovery and analysis across three of the world's largest machine actionable folklore collections (Dutch, Danish and German), thereby offering researchers an unprecedented opportunity to discover and interrogate spatial, temporal and network patterns both within and across the target corpora. By making search / discovery paths addressable, ISEBEL promotes the sharing of (interim) results, while the export of results offers support for additional offline analysis. A suite of tools designed specifically for computational folkloristics supports rich analytical engagement with the material, ranging from geo-temporal and network visualizations to statistical summaries of folklore data. Our work provides much-needed state-of-the-art infrastructure architecture for the ongoing development of folklore archives which protect, preserve and make accessible folklore as cultural heritage worldwide, while illustrating the power of computational methods for the analysis of folk traditions. The extensibility of ISEBEL, as part of its core architecture, ensures that other collections and additional tools can be easily added in the future.

Folklore is particularly suited to large scale data-driven analysis, as the field is predicated on the study of hundreds or thousands of variants of folk expressions. Since folklore can be conceptualized as the circulation of traditional expressive forms on and across social networks embedded in time and space, research tools that can handle this complexity (millions of variants told by hundreds of thousands of individuals in tens of thousands of places to hundreds of collectors) are those that provide the greatest benefit to researchers and the general public.

Folklore is generated and circulates at multiple scales, while understanding folklore necessitates a multiresolution approach, spanning the continuum from traditional close reading to distant reading, and back. Consequently, ISEBEL is conceptualized as an instantiation of Katy Börner's plug-and-play macroscope (Börner 2011), where the macroscope "provide[s] a 'vision of the whole,' helping us 'synthesize' the related elements and detect patterns, trends, and outliers while granting access to myriad details. Rather than make things larger or smaller, macroscopes let us observe what is at once too great, slow, or complex for the human eye and mind to notice and comprehend."

ISEBEL takes advantage of existing work on computational folkloristics by deploying and extending existing open source tools on a series of integrated data collections, thereby moving a giant step toward realizing Börner's notion of the plug-and-play nature of the macroscope. Notably, our data collections represent the most comprehensive machine actionable folklore collections in the world, comprising hundreds of thousands of records, and the repertoires of thousands of storytellers, spanning well over a century.

1.a Problem Overview

A desideratum of folklore research has, ever since the inception of the field, been the ability to work comparatively across cultural and linguistic boundaries, an idea that informed the large indexing projects of the late nineteenth and early twentieth centuries (Uther 2004; Thompson

1966; Christiansen 1977). More regional approaches to indexing are always done with an eye toward this type of broad research (Klintberg 2010). While some of the early indexing efforts were clearly predicated on a desire to understand developmental trajectories of stories as well as the origins of various beliefs (Krohn 1926), a great deal of recent research has been interested in situating belief in place and time, and using this as a means for exploring how people use stories as a means for the ongoing negotiation of cultural ideology (Tangherlini 1994). Both approaches – and many lying on the continuum between them – rely on good collections and even better finding aids.

National archives have by design been largely available for folklore research since their establishment. Yet, the exigencies of local needs balanced with the limited availability of archivist time to index and classify materials necessarily meant that the archives supported local questions first, and responded to more broad scale considerations second. The development of international indices for aspects of traditional expressive culture ensured at least, on some level, a degree of "interoperability" (Uther 2004). A researcher interested in stories about a particular topic or a particular ATU number could write to a series of archives and receive fairly comprehensive responses, although the responses would, more often than not, be in the format of the local archive, and might require a degree of knowledge of a particular collection's structure to understand those responses. Similarly, there was no obvious manner in which to evaluate the thoroughness of the reply – did it represent the results of a comprehensive search or was it the result of a sampling of the archive's holdings? Often, there was little support in this type of search for "query expansion", where the original research terms could be expanded, based on local archival knowledge or other forms of domain knowledge, to include other related terms.

The digital revolution held the promise of making archival research both more efficient and more thorough. For the digital revolution to be fully realized, however, it required a degree of technical knowledge and digitization resources such as high-quality high-speed scanners that were frequently out of reach of all but the most privileged of archives. Yet nearly all archivists recognized that, if archival records were digitized and stored with appropriate metadata, then a researcher would be able to access, at least in theory, the records needed to support her work by generating a search over the archive – the results would be returned much quicker than through the physical search of a paper or tape-based archive as described by Gunnell, or by requesting resources via the mail (Gunnell 2010). Even prior to the broad-based adoption of the internet in the late 1990s and the migration in the early twenty-first century of archival collections to databases queried through web-based search interfaces, the results of a search could be returned via email in a matter of hours or days, instead of weeks and months. Similarly, a researcher could expand queries quickly, learning from the results of very precise queries to generate searches that would substantially increase the recall over the archival collection (Singh et al. 2016).

A great deal of archival thought and folkloristic expertise ensured that the migration from wholly analog records (i.e. paper; photographic and moving images recorded on film; sound recorded on cylinders, vinyl and magnetic tape; material artifacts) to largely digital records proceeded much more quickly than if the archives had to start from scratch. The accessibility and low cost of well-supported internet infrastructure bundles, such as the open source Linux-Apache-MySQL-PHP (LAMP) stack that included robust solutions for relational databases and web-based querying, also ensured a degree of commonality across many folklore archives. Since these

archives generally consist of collections created by individual collectors or networks of collectors from interactions, either in person or through mailed surveys, with informants whose performances of informal expressive culture were the main targets of collection, one could formally represent the overall structure of these archives as a tripartite network of persons – places – things (Tangherlini & Broadwell 2014). In this model, persons can play multiple roles such as archivists/classifiers, collectors, or informants. Places can likewise have multiple roles, related to any of the persons, or to the things the persons perform (e.g. places mentioned in a legend), while things can be conceived of as covering any of the broad types of folklore stored in the archives. Importantly, this conceptualization of folklore – as informal cultural expressive forms circulating on and across social networks embedded in time and space – does not presuppose that folklore is the product of a deliberate collecting process, and can just as easily be applied to "born digital" folklore that may be archived asynchronously and indirectly.

The demands of the digital folklore archive, then, are closely aligned with the archival practices of the past century [fig. 1]. The more recent proliferation of metadata standards throughout the library and archival world aligns remarkably well with the pre-existing deep description of folklore materials in the various archives, and has allowed for the quick adaptation of digital approaches to folklore archiving that one encounters across most of Northern Europe, even in the face of severe budgetary constraints. This convergence on relatively similar systems, the ability to create meaningful queries that are not constrained by previous indexing regimes, and the ability to integrate search results with helpful user interfaces including various types of visualizations, has allowed for a rapid increase in the usability of the now-digital archive.

Figure 1: A view of the Wossidlo Archives.

The metadata for the Icelandic folklore collections of Jón Árnason, for instance, form the bedrock for the geographic navigation system of the *Sagnagrunnur* system developed by Gunnell and Dagsson, which can then be integrated with other folklore collections at the AMI (Gunnell 2010). Similar infrastructure has been used to support a geographically based user interface to search legends in Sweden and Norway through the *Sägenkartan* (Skott 2017). In the Netherlands, a longstanding effort of the Meertens Instituut has led to the development of perhaps the most comprehensive online folklore collection in Europe, with a clear and easy to use interface at the *Volksverhalenbank* (Meder et al. 2016). Fairly complex database models predicated on hypergraphs have allowed researchers at the Wossidlo Archive to create a series of research interfaces to explore the remarkable holdings of that archive, and to bring together otherwise disparate representations of a single story performance in *WossiDiA* (Bruder et al. 2015) [See Fig 2].

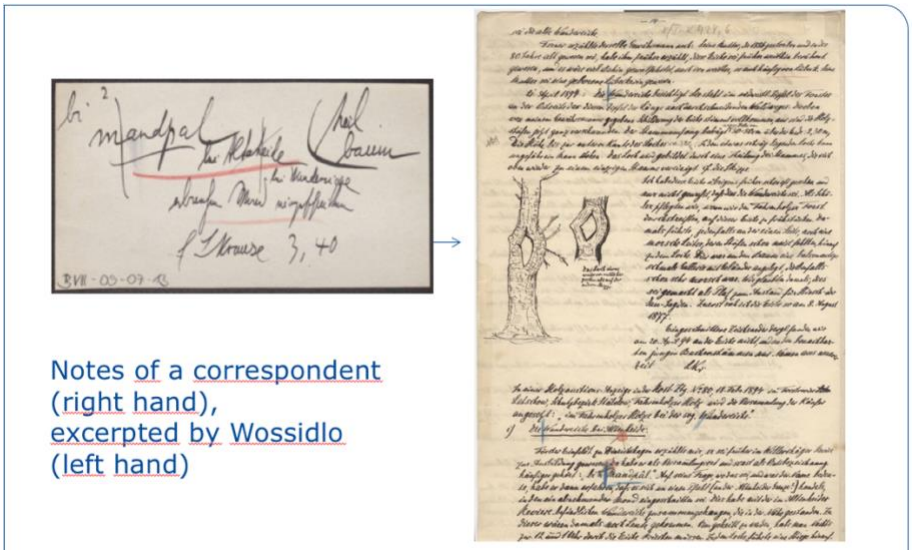


Figure 2: The various representations of a single story in the Wossidlo archive.

Danish folklore collected by Evald Tang Kristensen has formed the basis for a number of experiments in computational folkloristics, including various interfaces for user search such as *ETKspace* and *The Danish Folklore Nexus* (Tangherlini 2013a; Tangherlini & Broadwell 2014) as well as flexible network classifiers that leverage the power of existing classification systems, while taking advantage of the metadata associated with stories and storytellers in the people-places-things model of folklore noted above (Abello et al. 2012; Broadwell et al. 2017) [fig 3].

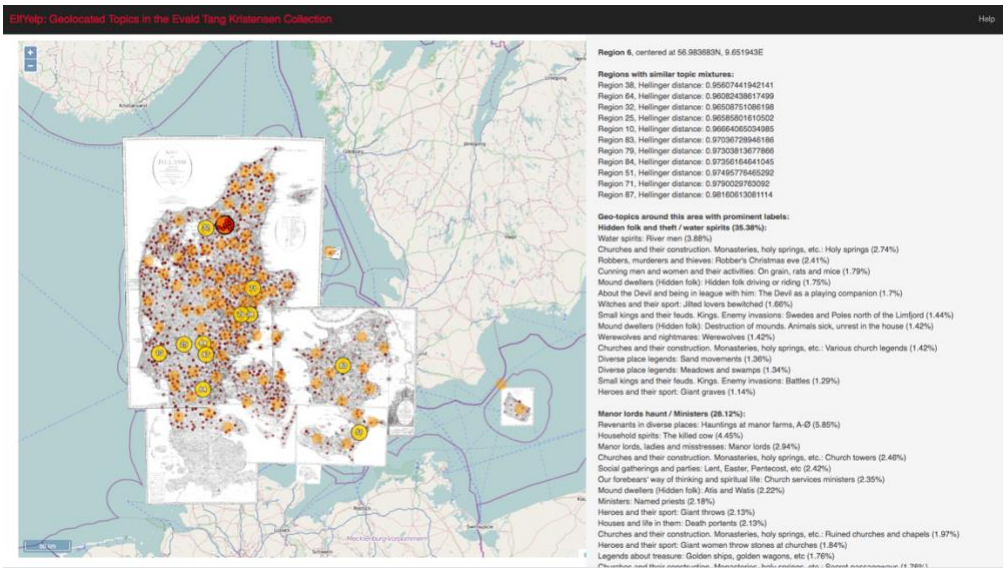


Figure 3: Examples of an analytic interface for the study of Danish folklore, here geolocated topic models, ElfYelp.

Other analytic approaches based on the digital representation of the Tang Kristensen folklore collection include multi-scale statistical representation of legend repertoires (Tangherlini 1994), a series of geographically based studies of distribution patterns (Tangherlini 2010; Broadwell & Tangherlini 2016; Broadwell & Tangherlini 2017), topic modeling (Tangherlini & Leonard 2013; Broadwell & Tangherlini 2015), text reuse (Broadwell et al. 2018), and the impact of transportation infrastructure on Tang Kristensen's collecting practices (Storm and Tangherlini 2018). While all of these studies have considerably altered not only the type of questions that one can address in folkloristics but also the manner in which one can address them, they are still largely constrained to a single archive, or a very small subset of archives (as in the case of the Swedish/Norwegian legend map noted above). In addition, nearly all of these archives restrict search to the language or languages represented by the records in their archival collections.

In short, research that crosses the boundaries of multiple archives still requires one to visit each archive individually – albeit virtually – and create archive-tailored searches. Results are often presented in a manner that precludes easy download and few, if any, of the archives provide an API for more automated search of the archival records. Furthermore, laws governing access can, in certain instances, make it impossible for international researchers to access sophisticated local search interfaces. Consequently, despite the convergence of folklore archives on relational databases and the broad acceptance of international metadata standards for representing the data, there are several major barriers to realizing the goal of search across multiple archival resources, irrespective of language, and to work with those results in sophisticated analytical environments, be they geographic information systems (GIS), text mining environments such as *Voyant* (Sinclair & Rockwell 2016), or simply as a series of documents subjected to the standard philological and close reading methods of folklore. While one can, given a certain degree of patience, some limited programming background, and a fairly broad linguistic competence, find results for a search on "ghosts" in northern Europe by visiting the various online archives mentioned above and capturing those results locally, it is nearly impossible to work with those results in a "macroscopic" fashion (Tangherlini 2013b). Börner reminds us that the macroscope

provides a "vision of the whole, helping us synthesize the related elements and detect patterns, trends, and outliers while granting access to myriad details. Rather than make things larger or smaller, macroscopes let us observe what is at once too great, slow, or complex for the human eye and mind to notice and comprehend" (Börner 2011: 60). To realize an international – or at least broadly regional – macroscopic approach to folklore collections, one needs to have several facilities built-in to a search engine: (i) the ability to search once and retrieve results from multiple archives; (ii) the ability to search in a single language, and receive high-confidence results from those searches across all of the collections, irrespective of the language of the target repositories; and (iii) the ability to work either in the search environment or locally with the returned results, ideally with various tools for visualization and statistical analysis (Ilyefalvi 2018). The international project ISEBEL (Intelligent Search Engine for Belief Legends) has addressed at least the first two of these challenges, and the lessons learned during the process of developing the infrastructure for multi-lingual search across disparate folklore archives can help chart the path forward for addressing the third challenge as well as other future projects (Meder 2018; Schmitt & Tangherlini 2018). [fig. 3b]

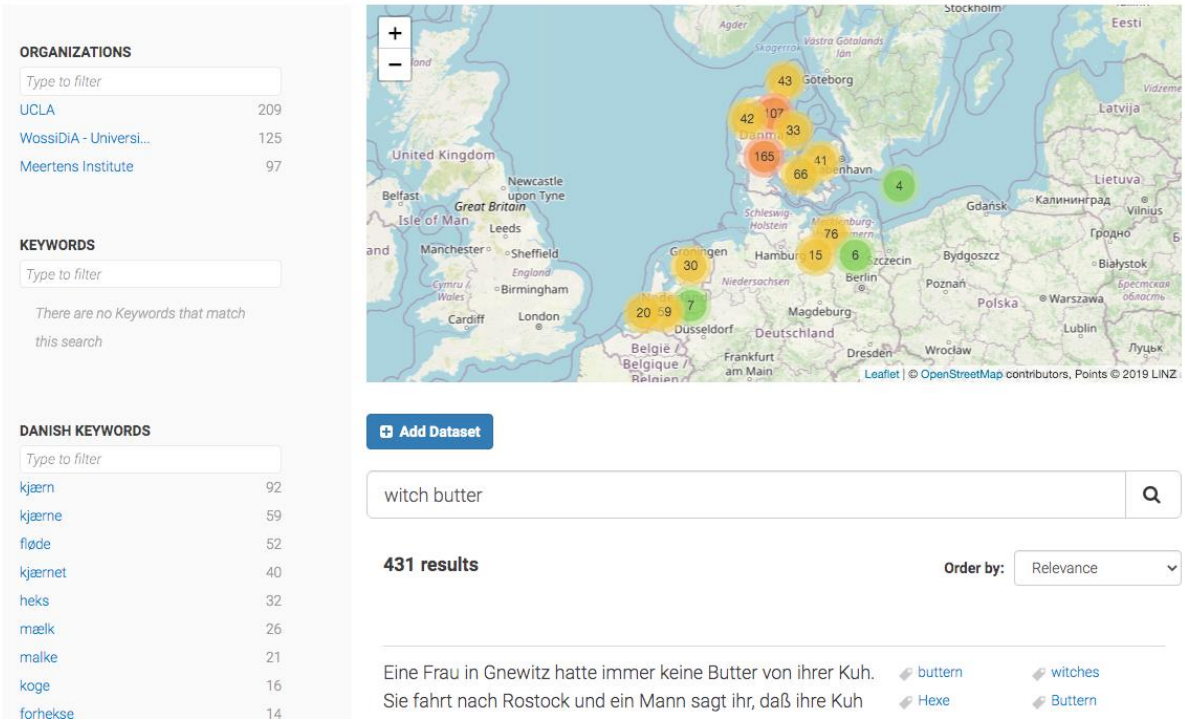


Figure 3b: A search in ISEBEL for "witch" and "butter" returns 431 results from all three archives.

2. Project Activities

Project activities were distributed across the three different institutions, each one taking a lead on a major component of the ISEBEL project, as well as series of local archive specific tasks. The Meertens Institute, because of the significant cost-sharing commitment of the KNAW and pre-existing infrastructure for hosting, maintenance and persistence took the lead on three core components of the project:

- The development of the main OAI-PMH harvester node and accompanying mySQL database for the storage of the harvested data
- The development of the CKAN-based GUI
- The coordination of the install scripts for OAI-PMH clients

The UCLA/Berkeley/Stanford team took the lead on two key components of the project:

- The development of the NMT-based translation infrastructure to support multi-lingual search
- The development of the domain key-term data for inclusion in the NMT translation pipeline

The Uni-Rostock team, because of their expertise in computer science, took the lead on:

- The development of the PowerGraph representation of the data for complex search, particularly for the challenging data representation at the foundations of the Wossidlo archive.
- The development of a clear work flow for the addition of new materials as well as semi-supervised translation of dialectical or otherwise unusual diction in stories.

The three teams worked extremely closely over the course of the grant period to develop methods for data representation, data harvesting and, starting in year two, multi-lingual search and the useful display of search results.

2.1 Developing the OAI-PMH Nodes for ISEBEL

Over the past three years, researchers associated with collections representing Denmark, Germany and the Netherlands have, under the umbrella of the aforementioned “Intelligent Search Engine for Belief Legends” (ISEBEL) project, explored ways in which to address this type of information need, recognizing that a series of disjoint keyword searches across the various collections, while a reasonable strategy, would not be likely to provide a coherent dataset for study.

The ISEBEL group quickly settled on implementing OAI-PMH (open archive initiative protocol for metadata harvesting) as the main method for accessing the underlying assets that each archive provider had deemed appropriate for inclusion in the search engine, and that could conform to local archival requirements for data sharing.

An important first step was agreement on a minimal data representation, expressed as an XML schema. Such a schema would facilitate integrating the collections on a “middle level” tuned to the search requirements of researchers interested in legends. While describing the inner workings of the entire ISEBEL system is beyond the scope of this white paper, the XML schema allows the collections to be accessed via the Open Archive Initiative Protocol for Metadata Harvesting (OAI-PMH) in a standardized form, and to be indexed in a manner that facilitates rapid search over the concatenated collections (Van de Sompel et al. 2004). The concatenated collections can

then be indexed and act as the target for search. This approach solves several problems at once: first, it obviates the need of researchers to visit multiple archives and create individual search strategies for each collection and, second, it allows each collection local control over what materials they offer for this type of concatenated search. This latter consideration is particularly important when collections have limitations on which materials researchers from outside of the archive are allowed to access.

The Danish OAI-PMH node, for example, is a clone of Open Culture Consulting's Simple OAI-PMH 2.0 Data Provider. The only modifications for the Danish Folklore provider are the metadata schema definitions in oai2config.php and the addition of the GenerateOAIFilesI2.py script, which exports metadata from the Danish Folklore MySQL database into XML files containing metadata records in the ISEBEL v2 metadata schema, which are then served via the OAI-PMH provider.

Each archive had slightly different approaches to data description and representation and, consequently, the group spent significant effort to develop a flexible yet overarching XML schema that represented a minimal set of data necessary for ISEBEL to operate as expected. [fig. 4]

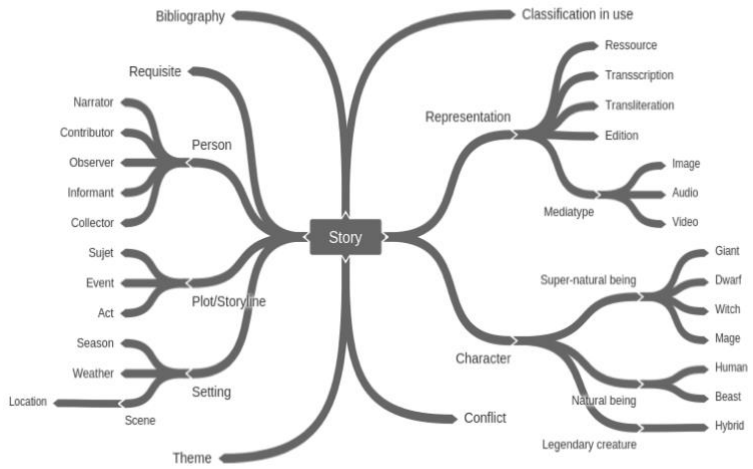
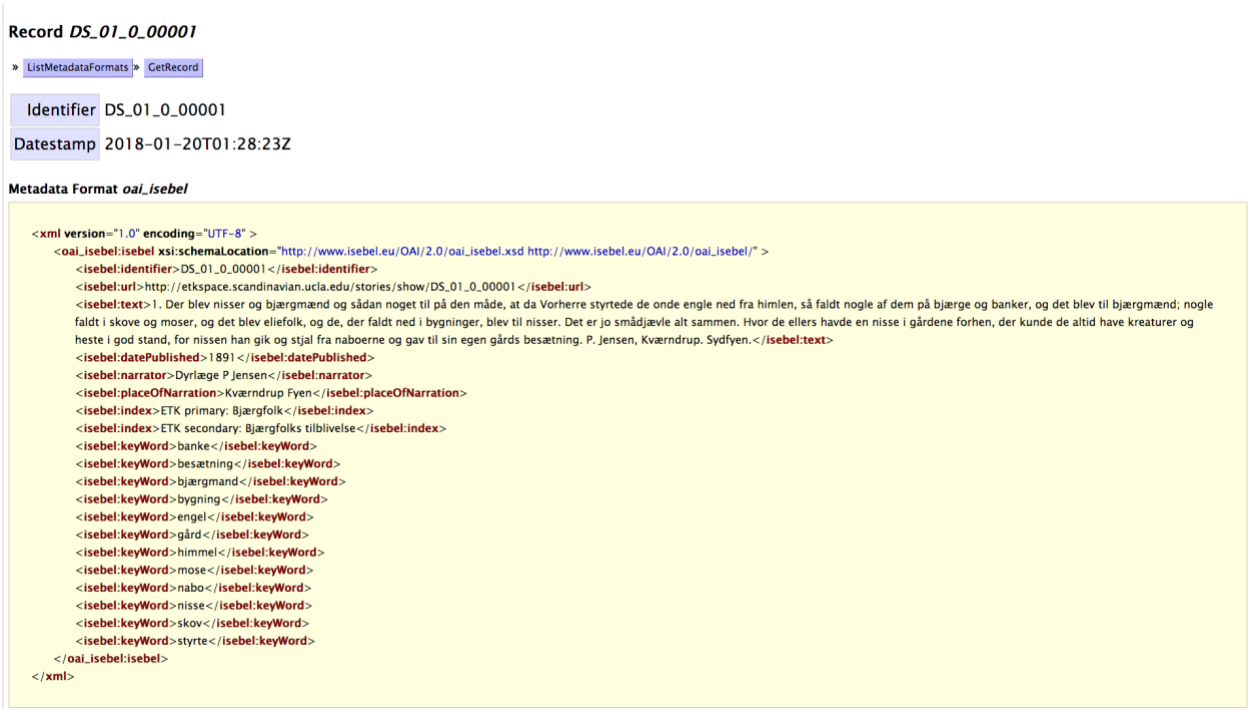


Figure 4: General overview of the components of a story in ISEBEL.

As a consequence of these considerations, considerable time was spent devising a flexible, yet easy-to-implement XML schema, allowing additional archives to easily join the initiative. At the same time, the schema needed to include facilities for supporting the multi-lingual search strategy that is a core functionality of the project. [fig. 5]



Record *DS_01_0_00001*

» ListMetadataFormats » GetRecord

Identifier	DS_01_0_00001
Datestamp	2018-01-20T01:28:23Z

Metadata Format *oai_isebel*

```
<xml version="1.0" encoding="UTF-8" >
  <oai_isebel:isebel xsi:schemaLocation="http://www.isebel.eu/OAI/2.0/oai_isebel.xsd http://www.isebel.eu/OAI/2.0/oai_isebel/" >
    <isebel:identifier>DS_01_0_00001</isebel:identifier>
    <isebel:url>http://etkspace.scandinavian.ucla.edu/stories/show/DS_01_0_00001</isebel:url>
    <isebel:text>1. Der blev nisser og bjærgmænd og sådan noget til på den måde, at da Vorherre styrtede de onde engle ned fra himlen, så faldt nogle af dem på bjærge og banker, og det blev til bjærgmænd; nogle
    faldt i skove og moser, og det blev eliefolk, og de, der faldt ned i bygninger, blev til nisser. Det er jo smådjævle alt sammen. Hvor de ellers havde en nisse i gårdene forhen, der kunde de altid have kreaturer og
    heste i god stand, for nissen han gik og stjal fra naboerne og gav til sin egen gårds besætning. P. Jensen, Kværndrup. Sydfyen.</isebel:text>
    <isebel:datePublished>1891</isebel:datePublished>
    <isebel:narrator>Dyrlæge P Jensen</isebel:narrator>
    <isebel:placeOfNarration>Kværndrup Fyen</isebel:placeOfNarration>
    <isebel:index>ETK primary: Bjærgfolk</isebel:index>
    <isebel:index>ETK secondary: Bjærgfolks tilblivelse</isebel:index>
    <isebel:keyWord>banke</isebel:keyWord>
    <isebel:keyWord>besætning</isebel:keyWord>
    <isebel:keyWord>bjærgmand</isebel:keyWord>
    <isebel:keyWord>bygning</isebel:keyWord>
    <isebel:keyWord>engel</isebel:keyWord>
    <isebel:keyWord>gård</isebel:keyWord>
    <isebel:keyWord>himmel</isebel:keyWord>
    <isebel:keyWord>mose</isebel:keyWord>
    <isebel:keyWord>nabo</isebel:keyWord>
    <isebel:keyWord>nisse</isebel:keyWord>
    <isebel:keyWord>skov</isebel:keyWord>
    <isebel:keyWord>styrte</isebel:keyWord>
  </oai_isebel:isebel>
</xml>
```

Figure 5: An example of a story, harvested by the OAI-PMH harvesting node at Meertens from the Danish folklore archive at UCLA.

The complete schema for ISEBEL can be found at: https://github.com/vicding-mi/isebel-schema/blob/master/xsd_test/isebel2.xsd

2.2 The PowerGraph Model, GraphMining and the Overall Architecture of ISEBEL

The ISEBEL system, after application of the OAI-PMH harvesting, creates an intermediate layer on which several additional systems operate. [fig 6]. These include search interface, which implements a custom version of the open source CKAN data portal platform.

In addition, the intermediate layer exposes the concatenated collection to two projects that are implemented in early alpha- or beta- versions. The first is the existing PowerGraph system that relies on hypergraph modeling of the complex, heterogeneous data in the concatenated collection. This approach was already tested and implemented for the Wossidlo archive as WossiDiA: https://www.wossidia.de/. The second is a GraphMining system currently in alpha development that allows for fast, yet complex, mining of the data.

Since the original archives all were based on different data schema, the need for a coherent minimal schema as discussed above allows for the harvested collection to be available for multiple search, mining, and analytic layers.

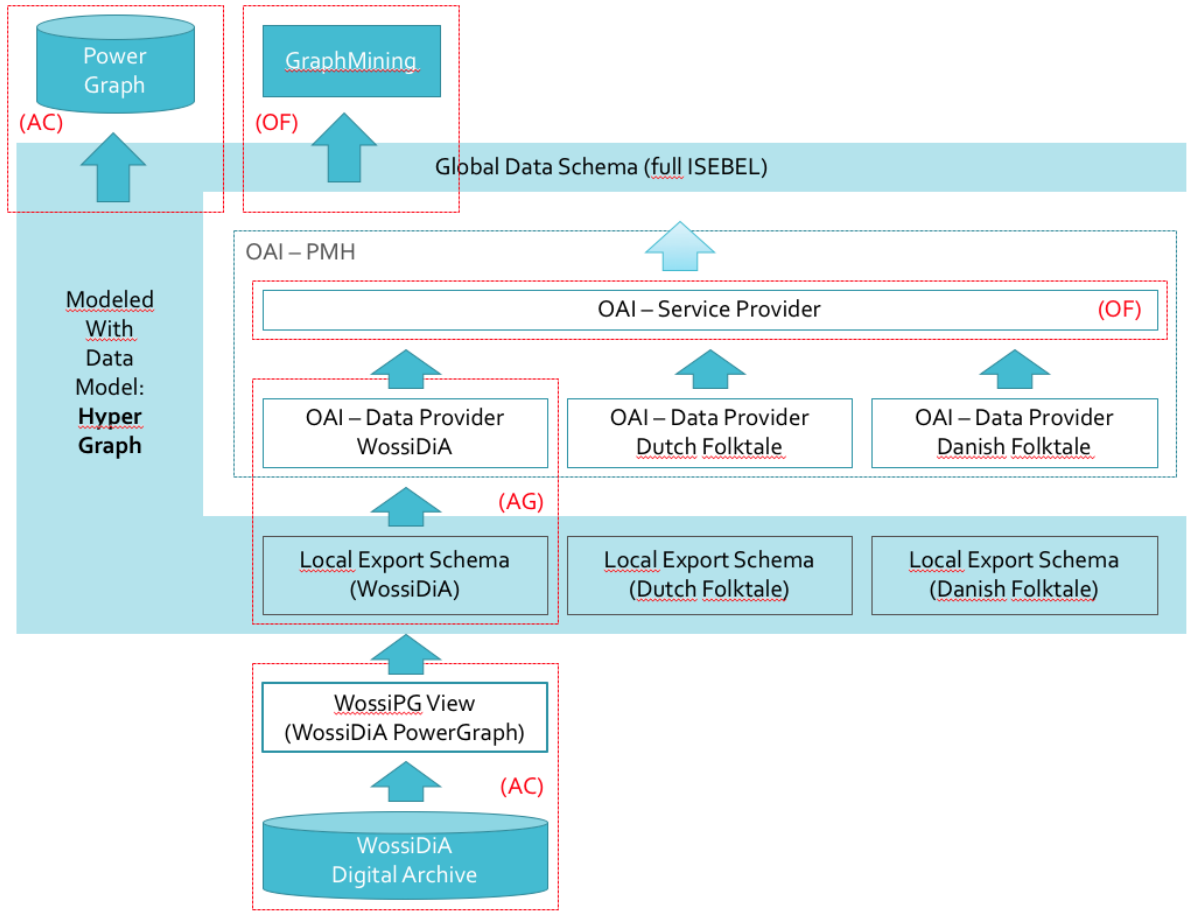


Figure 6: The architecture of the ISEBEL system including the representation of the individual archives harvested through the OAI-PMH provider-harvester nodes.

2.3 Multi-lingual Search for Belief Legends

Search across collections of multi-lingual data poses a series of well-known challenges for which various strategies have been proposed (Aula 2009). In the study of folk legend, an example of an information need could be posed as follows: “I am interested in ghosts, and legends about ghosts, from folklore collections in German, Dutch, and Danish.” The problem presupposes the existence of searchable collections for each of the target languages, and some way to access those collections. Earlier, a search might require an exchange of letters or physical presence at each of the archives, as well as time spent learning about the archival practices in each of the countries. It would similarly presuppose linguistic competence in each of the target languages.

While the OAI-PMH structure solves several major barriers to developing research collections that cross both national and linguistic borders, it does not immediately address one of the most vexing challenges to working across languages, namely searching in one language and receiving appropriate results from all of the languages in the integrated collection. Indeed, it was this problem that motivated the earlier indexing work prevalent in folkloristics in the mid twentieth

century. Search across collections of multi-lingual data poses a series of well-known challenges for which various strategies have been proposed, including translating the search string into the target languages at the time of search, running a search string in one language against translations of the target data which have been translated previously either into a common language or into the search language, limiting the vocabulary allowed in a search to terms in a multilingual thesaurus, or some combination of those strategies along with the use of pre-existing indices, such as the ML index (Trojahn et al. 2014; Christiansen 1977).

Each strategy has its distinct advantages, as well as fairly significant disadvantages. Translating all of the legends in one collection into the languages of all the other collections is time consuming (computationally expensive) and, depending on the quality of the translation required, quite likely impossible. Given the potentially very large number of languages, including dialects and other low resource languages that may be found in folklore archives, the scale of the problem could very quickly become intractable. Translating search strings into the target languages might be possible, but given challenges in both syntax and semantics (i.e. the search program would need to guess at the intention of the searcher, and translate that intention into multiple languages) unlikely to succeed in a highly specialized domain such as folklore. To sidestep these problems, one could limit searches to some sort of a controlled vocabulary so that the translations of those terms were of high quality and of clear relevance to each of the collections. This approach necessarily channels the searcher into constrained searches that presuppose a range of information needs. Because of the scale of the collections, even for a modest pilot such as ISEBEL, hand translation is impossible, and some degree of machine translation is inevitable (Table 1). The goal of any such machine translation, then, is to provide the best possible representation of a language for the lowest computational cost.

Collection	Total Number of records in ISEBEL	Languages in collection	Total number of keywords
Verhalenbank (NL)	26866	Dutch, Frisian, numerous dialects	28820
WossiDia (DE)	11622	High German, Low German, numerous dialects	2079
ETKspace (DK)	31086	Danish, Jutlandic dialects	1375

An overview of the size of the ISEBEL collections, including total number of records, languages represented in the collection, and total number of keywords associated with the collection.

2.4 ISEBEL's Hybrid Approach to Multilingual Search

Given the lack of resources for many of the languages represented by ISEBEL – a characteristic of many folklore collections – we devised a multipart strategy to facilitate multilingual search. The goal was to present the researcher with multiple avenues for search, including faceted search, where pre-existing collection-based indices could be turned on or off as appropriate, and keyword search, where the existing keyword indices for each corpus could be used either as a primary or secondary search parameter. Importantly, the ISEBEL XML schema explicitly captures this metadata (collection-based keywords and indices) during the harvesting process. Even with this limited data, one could in theory mimic a multilingual search across all of the collections by simply entering terms for the query in each of the target languages. A search on the terms for "ghost" in German, Dutch and Danish – "Gespenst OR spook OR spøgelse –

properly returns hits from each of the collections but requires the user to do her own multi-lingual query expansion. While words for ghosts are easily generated through dictionary look-up, more unusual terms, such as the Danish *lindorm* or the Dutch *kabouter* would likely not be found in most dictionaries. These two problems in multilingual search for folklore archives can be summed up as: (i) the need for a common search language and (ii) the need for a multilingual thesaurus of domain specific terms. In the ISEBEL project, we address these two issues separately, and then combine the results in the concatenated database records that form the target for search.

Generating domain specific terms for each corpus was a fairly straightforward task. We designated a domain expert for each target collection, who was then tasked with creating a list of terms that were likely to appear in the text of a legend but unlikely to be translated by a machine translation system. To illustrate this, the Danish term *nisse* is widely used in the Evald Tang Kristensen collection, appearing in 585 stories, but improperly translated by most machine translation systems as "elf" rather than the more appropriate "house elf" or "brownie." After each corpus expert had created a list, they were then asked to translate these lists into English, with multiple translations stored in separate rows. Similarly, the Danish term *lygtemand* has multiple possible English entries including "fen fire", "jack-o'-lantern", and "sprite" to name but three, while the terms *bjærgfolk* and *højfolk* both resolve to the English term, "hidden folk" while *højfolk* can also be translated as "mound folk". Once all of the lists were concatenated, unaligned terms from each of the other languages or dialects were translated into each of the collection languages if a corresponding term existed. These terms were then added to the indexing for each of the stories such that a Danish story that included the term *bjærgfolk* was indexed with the English term, "hidden folk." A Danish story with the term *højfolk* would similarly include the "hidden folk" indexing, and a search on "hidden folk" would subsequently return both stories. Since German and Dutch stories are indexed via the combined domain specific term list, this search would also return records from each of those collections for *Erdmännchen, Unterirdische, aardmannetje, verborgen volkje Ierdmännken, ierdmantsje,* and *forburgen folkje* (See Table 2). The incorporation of the domain specific terms greatly increases the accuracy of search, particularly given the specialized vocabulary of storytellers, while also simplifying the creation of a comprehensive search string. Due to the manner in which these additional terms are included in the search, the domain specific terms list can be updated easily, thereby increasing the accuracy of searches. Because of this feature, any archive that joins the ISEBEL consortium in the future will be required to generate a list of domain specific terms with their English language equivalents.

German	Dutch	Danish	English	Low German	Frisian
Erdmännchen	aardmannetje	bjærgmænd	hidden folk	Ierdmännken	ierdmantsje
Hexe	toverkol	heks	witch	Hex	tsjoenster
Wassermann	meerman	havmand	merman	Wasserman	mearman

Some examples of the domain specific key terms from the ISEBEL alignment table.

Along with keyword indices and domain specific terms, ISEBEL makes use of neural machine translation (NMT) to produce "dirty translations" of all the stories in the harvested corpuses. These dirty translations are then indexed and exposed to the search engine. NMT, which represents the current state of the art in machine translation, received a considerable boost in the

past two years with the development of what is known as the "Transformer" architecture. This architecture was developed specifically for machine translation tasks, where an input sequence must be accurately transformed into an output sequence. While earlier models of machine translation tended to be built on rules-based models of translation between two languages (direct translation) and, later, on abstraction based models (interlingua translation and transfer systems), current machine translation systems make use of very large corpuses of paired sentences (source→target) for training purposes and take advantage of the power of multi-layered neural networks to learn, in an unsupervised manner, high dimensional feature spaces (Hutchins 1995). These new models rely on a mathematical approximation of "attention" that encodes how source and target words are likely to be associated with each other across different contexts as a means for training the encoders and decoders that comprise the Transformer (Sutskever et al. 2014; Cho et al. 2014; Bahdanau et al. 2014).

For ISEBEL, we implement a modified version of the OpenNMT-py BPE (byte-pair encoding) pipeline, a "multi-headed attention" model that, along with the encoding of the input, greatly speeds the training of the model and the accuracy of its output (Vaswani et al. 2017; Sennrich et al. 2105). The BPE or "word piece" component is a powerful refinement to Transformer training practices that uses byte pairs (roughly equivalent to letter pairs) as its primary token vocabulary, rather than full words. This has the effect of shrinking the overall unique token count of the corpus, greatly reducing the amount of GPU RAM required. It also tends to produce more accurate translations. Part of this increased accuracy, particularly for the tasks related to creating indexable translations of legends, is that the BPE pipeline typically does a better job of recognizing untranslatable terms (e.g., proper names or supernatural creatures) and leaving them unaltered in the final translated output, whereas full word-based systems may drop these terms from the vocabulary entirely as a memory-saving measure due to their rarity. This latter attribute of word-based systems often makes it quite a chore to recover the original forms of dropped words and retain them in the output translations. BPE also tends to make the Transformer system better able to encode how minor orthographic and morphological changes might modify the "meaning" (in a very narrow, computational sense) of a word.

Currently, every record in ISEBEL is translated into English via the byte-pair encoding Transformer approach described above. Training these translation models requires large "parallel" corpuses of equivalent sentence pairs in the source and target languages, most of which we obtained from the Open Parallel Corpus (OPUS) project (Tiedemann 2016). The availability of such open text sets for moderately and even under-resourced languages greatly accelerates the process of creating rich source→target training sets, and provides a clear path forward for the addition of future folklore collections into the ISEBEL framework. As of this writing, fully pretrained models for translating to and from some of the most highly resourced joint translation counterparts of English (French, German, Chinese) are available for download, as are rudimentary trained Transformer models based on many of the languages and paired texts from the OPUS project. It is not always easy to adopt models others have trained, nor is this necessarily desirable when working with regional and historical corpora. Such models may be quite accurate for translating modern prose or speech as they have been trained on extremely large contemporary news, web, literature, and multimedia corpuses, but they often fail to perform adequately when translating, for example, a nineteenth century regional dialect. Simply training a model with the historical texts alone also is not effective due to their much smaller

number. Often the best method – which we adopt for ISEBEL – is to train a model with numerous contemporary parallel corpora such as those from OPUS to establish the model's linguistic "backbone," and to augment this with additional materials from more period appropriate, publicly available translation resources such as the Bible, or even manually translated texts from our own corpuses. The Transformer architecture and NMT models' voracious appetite for training data also make it possible and frequently desirable to train a single English-targeted model with corpuses from multiple closely related source languages, such as Danish and Swedish; the resulting model is often more accurate when translating from either source language to English than models that are trained separately.

When possible, we refine our models by comparing the accuracy of the NMT translations with hand generated "gold standard" translations – though not the same texts used to train the model – and, after discovering areas of systematic failure, update the model to solve these problems (Table 3). It was this type of an evaluative approach that led us to adopt the BPE method and to devise and incorporate the "domain specific term" search augmentation workflow described above, as we discovered a series of terms that the initial Transformer consistently failed to translate.

Original Story Phrase	Prior translation (Transformer)	Current Translation (Byte pair encoding)	Gold Standard
Så siger den fattige Mand, at han havde en Karl, der rimeligvis kunde begrave hende sådan, at hun skulde ikke komme igjen.	Then the poor man says that he had a farmhand who could bury her so that she wouldn't come again.	Then the poor man says he had a farmhand who could reasonably bury her so that she shouldn't come again	Then the poor man says that he had a farmhand who probably could bury her so that she wouldn't come back.
Bjærgfolkene i den bakke havde også undertiden deres linned ude.	The <unk> in that hill also sometimes had their <unk> out.	The mound men in that hill also sometimes had their linen out.	The mound dwellers in that mound sometimes put their linen out.
Heksemesteren havde sagt ham, hvordan det vilde gå til.	<unk> had told him how things would go	The witch master had told him how it would go	The witch master had told him what would happen
Men en gang var bjærgfolkene, der boede i højen, blevne fornærmede på dem	But once <unk>, which lived in <unk>, was <unk> upon them	But once the mound dwellers lived in the mound, they were insulted	But once the mound folk who lived in the mound were offended by them
En karl traf en gang et par snoge, der spøgte, og slog da den ene ihjel.	A farmhand once took a couple of <unk> that <unk> and then killed one	A farmhand once took a couple of songs that ghost, and killed one	A farmhand once met a pair of grass snakes that were slithering about, and he killed one of them

Examples of the various "dirty" translations of Danish phrases from a small selection of legends. In the original Transformer, untranslated words were replaced with <unk>. The quality of the translations using byte-pair encoding is significantly better than the earlier implementation. Incorporating the domain specific keywords provides even better search capabilities.

Initially, we had attempted to create a n_i x n_j translation system, where each language corpus was translated into the language of each of the other corpuses, but we soon ran into several performance barriers. Although we had access to fairly robust computing resources, the training

of the language models took significant time even with the performance gains offered by multi-headed Transformers with byte-pair encoding. Also, because there are few resources for translating between many low resource languages such as Danish and Frisian, the accuracy of the translations was quite low. In contrast, there are many more parallel corpuses and even pre-trained models for translating from many of the languages in our archives to English. By using English as the target translation and search language, we were able to sidestep several of these considerable performance barriers, while providing the facility for researchers to search in a single language and retrieve results from many languages.

As the grant was reaching a conclusion, we began exploring the possibility of taking advantage of multi-lingual embedding spaces, an approach that might be able to help address some of the problems associated with terms that are difficult to translate from one language to another, and that might provide some support for n_i x n_j translation, particularly in the case of very closely related languages, such as the Nordic languages. [fig. 7]

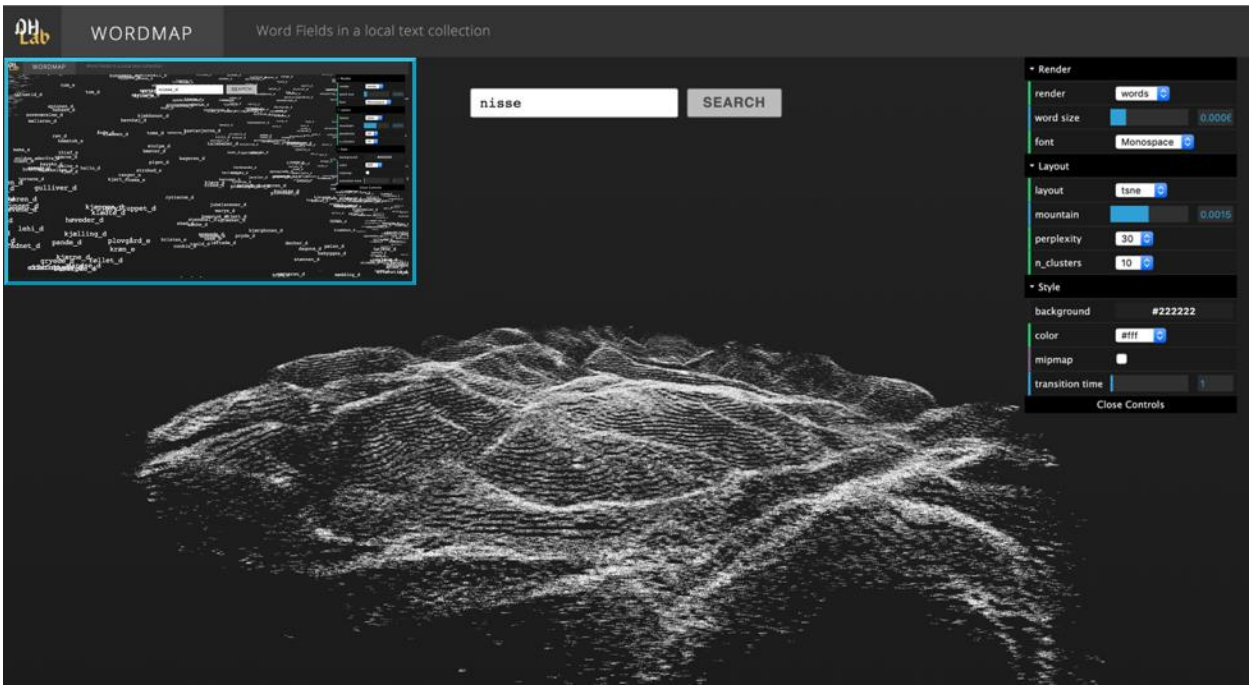


Figure 7: A multilingual embedding space, here with Danish and English. Inset, a search on the Danish word “nisse” and its multilingual neighbourhood.

3. Audiences

ISEBEL is focused on a broad range of audiences. The Meertens Institute already reports an increased interest in their already highly utilized online web portal for folk narrative that includes the subset of belief legends provided through their OAI-PMH harvester to ISEBEL. For the Meertens Institute, audiences include K-12 students, university students, the general public, as well as researchers in ethnohistory, literature, dialectology and cultural studies. Similar audiences are also reported for the Wossidlo Archive and the University of California collection of Danish folklore. This latter collection is particularly targeted toward undergraduate students, as well as the broader research community.

ISEBEL has drawn significantly broader audiences to the search engine, particularly those interested in the distribution of motifs across a broad swath of Northern Europe. As the Nordic collections become incorporated into ISEBEL, there is little doubt that the search engine will be able to service the information needs of many different groups from broad linguistic backgrounds.

Since ISEBEL has only gone "live" in the past few months, the administrators at Meertens Institute have yet to provide the project with detailed visitor numbers, or geographic spread. Later work would focus on user surveys so that we could understand not only who the visitors are, and their reasons for visiting ISEBEL, but also their paths through the collection, and the degree to which multilingual search across multiple archival resources has supported their research or provided substantive and meaningful answers to their questions.

4. Evaluation

The two main parts of the ISEBEL system that we were able to evaluate in a consistent manner were the accuracy of the OAI-PMH harvesting, and the usefulness of the multilingual NMT-based machine translations. Other evaluations, such as usability surveys focused on the graphical user interface were beyond the scope of this study.

For the OAI-PMH harvesting, we were able to check the success of harvesting through a comparison of the source stories and their encoding and the target version of the harvested stories. Although the preliminary harvesting was "noisy" and, in the case of the Wossidlo archive, missed entire tranches of stories, ongoing debugging and fine tuning of the system—along with the local implementation(s) of the XML schema—resulted in nearly error-free harvesting of all the provided data from each of the data providers. The lessons learned during this process are now part of our instructions (see the github repositories) for new archives interested in joining the ISEBEL consortium.

For the NMT-based machine translations, the evaluation of the translations was twofold. First, BLEU scores were calculated on the machine translations compared to reference translations created by domain experts. As noted above, the models we deployed were eventually able to create search target translations of such high quality that our current modifications to the CKAN GUI will allow end-users to expose the "dirty" NMT translation along with the text of the story in the original language.

5. Continuation of the Project

There are numerous initiatives focused on expanding the scope of ISEBEL so that it incorporates collections of belief legends from archives throughout the world. There are three main initiatives, one focused on Northern Europe, predominantly Scandinavia (Nordic Folklore Macroscope, now NordISEBEL), Finland and Ireland; a second group focused on Hungarian tradition archives; and preliminary explorations of integration with Greek tradition archives. Additional interest has come from folklore archives in East Asia, Africa, and South America.

5.a NordISEBEL

The Nordic Working Group in Computational Folklore proposes to develop an advanced digital environment that will align the rich folklore resources of the Nordic region. Such an environment will let users explore the diversity of oral traditions and belief practices across the region while offering researchers from many disciplines sophisticated access to these collections. The unifying principle of the project is the "folklore macroscope", a model that captures the complexities of folk culture, allowing users to move between all scales of analysis, from the micro-scale of close reading to the macro-scale of pattern discovery. Our Nordic folklore macroscope will help users from all walks of life explore Nordic culture in an immersive online context that encourages exploration and supports discovery.

Initially, the project will focus on integrating existing best-in-class projects from Denmark, Iceland and Sweden, while also providing a clear infrastructure to extend the work to contemporary ethnographic projects and the ongoing digitization of historical archives. Importantly, the project will focus on interoperability with international projects, such as the German-Dutch-American collaborative project, ISEBEL (Intelligent Search Engine for Belief Legends), and follow the practices of European digital humanities infrastructure projects, such as CLARIN.

As a first step, the working group will integrate the digital legend collections currently housed in the Danish Folklore Nexus (UCLA), Sagnagrunnur (Univ. of Iceland), and Sägenkartan (Institutet för språk och folkminnen). This work will make use of existing Open Archive Initiative harvesting protocols to generate a comprehensive intermediary collection that can be exposed to various search protocols, such as those integrated in the Elastic stack that drives Sagnagrunnur and Sägenkartan, or the analytic tools developed as part of the Danish Folklore Nexus. We have already implemented a pilot project allowing the integration of Sagnagrunnur data in ISEBEL.

Data harvesting and integration builds directly on the work of ISEBEL; as a result, once implemented, the Nordic datasets will be interoperable with the infrastructure and tools built through that project as well. Importantly, since all of the Nordic projects already rely on nearly identical data schema and database software, integrating these collections can happen quite rapidly, allowing us to move on to other key aspects of this proposal: (1) advanced search and visualization including multi-lingual query expansion, and advanced geographic visualizations; (2) machine learning for flexible classification; and (3) data curation for historical archives and contemporary collections tuned to the specific requirements of data access and protection governing each of the collections.

5.b Persistence

The Meertens Institute in the Netherlands has agreed to host and maintain the ISEBEL site for at least 10 years, through 2030. This commitment will allow the current infrastructure to continue to provide a central location for the harvesting node, as well as provide secure and robust web infrastructure capable of handling very high traffic. The Meertens infrastructure will also act as a secure data storage for the harvested data, as part of the archiving and preservation goals of the ISEBEL project.

Over the course of this period, ISEBEL members will continue to explore improvements and refinements to both the user interface. Updates and other maintenance issues for the ISEBEL harvesting, search and storage infrastructure will be supported by the Meertens Institute during this timeframe.

Members of the growing ISEBEL consortium are dedicated to identifying partner institutions as well as funding opportunities to continue to grow and improve the ISEBEL site.

6. Long Term Impact

We expect the long-term impact of ISEBEL to be significant, both in the field of folkloristics, but also in the field of multilingual search engines for concatenated archival collections. Already, the ISEBEL project has been the inspiration for a multi-year, multi-million NOK project at the Norwegian Royal Library. The project, SAMLA, is a five-year effort focused on the digitization of the various Norwegian tradition archives. As part of this effort, the goal is to create an OAI-PMH node, train a series of NMT models for various dialects of Norwegian, and allow these materials to be searched alongside the core ISEBEL archives, and the recently added Icelandic materials. Since we have developed a very clear workflow for additional archives to join the ISEBEL consortium, and have spent considerable time visiting with colleagues at numerous national archives of folklore, we expect in the not-to-distant future for visitors to ISEBEL to be able to search Catalan, Greek, Hungarian, Finnish, Latvian, Estonian, Icelandic and Norwegian materials along with the materials in German, Danish, Frisian and Dutch.

As the purview of collections included in ISEBEL grow, we expect follow on projects to incorporate media recordings (audio, video, images). By design, the ISEBEL schema is able to integrate these different types of media into the search. We have also experimented with integration with IIIF servers and have the facility for including deeplinks in the returned results. Consequently, ISEBEL has the potential to lead the way for future endeavors in multilingual search across multiple tradition archives.

7. Award Products

The ISEBEL project is now a live resource for the study of belief legend. All three archives can be searched in either the original language or in English at: https://search.isebel.eu

The various components and descriptions of the project are available through several github repositories.

The ISEBEL XML schema is available here: https://github.com/vicding-mi/isebel-schema/blob/master/xsd_test/isebel2.xsd

The various nodes for harvesting can be found here: https://git.informatik.uni-rostock.de/isebel/oai-pmh

The ISEBEL graph database, that lies at the basis of the PowerGraph model can be found here: https://git.informatik.uni-rostock.de/isebel/isebel-graph-database
Work on this part of the project is ongoing.

The team has presented numerous conferences and workshops, as well as published various articles in journals. For the US team, the most relevant publications are:

Peter M. Broadwell, Peter Leonard, and Timothy R. Tangherlini. 2018. "'Hvad der byggedes om dagen, blev revet ned om natten …': Word Sequence Repetition in Danish Legend Tradition." *Svenska landsmål och svenskt folkliv* 140(2017): 9-27.

Storm, Ida, and Timothy R. Tangherlini. 2018. " 'En temmelig lang fodtur': hGIS, Text Mining, and Folklore Collection in 19th Century Denmark." *Human IT* 14(2): 43-81.

Schmitt, Christoph and Timothy R. Tangherlini. 2018. "Folklore Archives Online. Zur Sichtbarmachung, Auswertbarkeit und Interoperabilität einer dänischen und einer nordostdeutschen Sammlung." *Jahrbuch für Europäische ethnologie 2018* 13(3): 181-204.

Timothy R. Tangherlini. 2019. "Modeling Anholt. Legend and Locality on a Nineteenth Century Island." In, *Former som formar: Musik, kulturarv, öar. Festskrift till Owe Ronström.* Edited by Camilla Asplund Ingemark, Carina Johansen, and Oscar Pripp. Uppsala: Uppsala University Press.

Peter M. Broadwell and Timothy R. Tangherlini. 2020. "Geist, geest, geast, spøgelse: Challenges for multilingual search in belief legend archives." *Arv: Nordic Yearbook of Folklore* 76: 7-28.

Timothy R. Tangherlini. 2020. The Dictionary of Jutlandic Folk Speech by Henning F. Feilberg. In, *Dictionaries as Sources of Folklore Data*. Ed. Jonathan Roper. Pp. 59-84.
Folklore Fellows' Communications 321.
Helsinki: Suomalainen Tiedeakatemia.

Timothy R. Tangherlini. 2021. "A Conspiracy of Witches." In, *Myth, Magic, and Memory in Early Scandinavian Narrative Culture. Studies in Honour of Stephen A. Mitchell*. Edited by Jürg Glauser and Pernille Hermann. Pp. 181-193. Acta Scandinavica 11.

Timothy R. Tangherlini. 2021. “Stumbling into folktales: Navigating the unusual collection of Nikolaj Chr. Christensen.” In, *Festschrift for Christoph Scmidt.* Edited by Petra Himstedt-Vaid. Rostock.

Publications from the other groups include:
Meder, Th. "ISEBEL: Intelligent Search Engine for Belief Legends." *Volkskunde* 119, no. 1 (2018): 87-89.

Meder, Theo. "The Technological Developments of the Dutch Folktale Database (1994–2016)." *Estudis de Literatura Oral Popular= Studies in Oral Folk Literature* 5 (2016): 45-69.

8. Bibliography

Abello, James, Peter Broadwell, and Timothy R. Tangherlini 2012: Computational folkloristics. *Communications of the ACM* 55(7).

Aula, Anne, and Melanie Kellar 2009: Multilingual search strategies. In: *CHI'09 Extended Abstracts on Human Factors in Computing Systems*.

Bächtold-Stäubli Hanns, Eduard Hoffmann-Krayer, and Lüdtke Gerhard 1927–42: *Handwörterbuch Des Deutschen Aberglaubens*. Berlin.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio 2014: Neural machine translation by jointly learning to align and translate. arXiv:1409.0473

Bendix, Regina 1997: *In Search of Authenticity : The Formation of Folklore Studies.* Madison.

Börner, Katy 2011: Plug-and-play macroscopes. *Communications of the ACM* 54(3).

Broadwell, Peter M., and Timothy R. Tangherlini 2015: ElfYelp: Geolocated topic models for pattern discovery in a large folklore corpus. In: *Proceedings of Digital Humanities 2014*.

Broadwell, Peter M., and Timothy R. Tangherlini 2017: Ghostscope: Conceptual mapping of supernatural phenomena in a large folklore corpus. In: *Maths Meets Myths: Quantitative Approaches to Ancient Narratives*. Basel.

Broadwell, Peter M., and Timothy R. Tangherlini 2016: WitchHunter: Tools for the geo-semantic exploration of a Danish Folklore Corpus. *Journal of American Folklore* 129(511).

Broadwell, Peter M., Peter Leonard, and Timothy R. Tangherlini 2018: "'Hvad der byggedes om dagen, blev revet ned om natten …': Word Sequence Repetition in Danish Legend Tradition." *Svenska landsmål och svenskt folkliv* 140.

Broadwell, Peter M., David Mimno, and Timothy R. Tangherlini 2017: The Tell-Tale Hat: Surfacing the Uncertainty in Folklore Classification. *Journal of Cultural Analytics*. doi: 10.7910/DVN/SYZ1PZ

Bruder, Ilvio, Holger Meyer, Alf-Christian Schering, and Christoph Schmitt 2015: Das Projekt WossiDiA: Digitalisierung des Wossidlo-Archivs. *Digitales Kulturerbe* 61.

Cho, Kyunghyun, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio 2014: On the properties of neural machine translation: Encoder-decoder approaches. arXiv:1409.1259

Christiansen, Palle Ove 1996: *A Manorial World : Lord, Peasants and Cultural Distinctions on a Danish Estate, 1750–1980*. Oslo.

Christiansen, Reidar Thoralf 1977: *The Migratory Legends*. International Folklore. New York.

Davies, Owen, and Willem de Blécourt 2004: *Witchcraft continued: Popular magic in modern Europe*. Manchester.

De Certeau, Michel 1984: *The Practice of Everyday Life*. Berkeley.

Dégh, Linda 2001: *Legend and Belief : Dialectics of a Folklore Genre*. Bloomington.

Dundes, Alan 1964: *The Morphology of North American Indian Folk Tales*. FF Communications 195. Helsinki.

Dundes, Alan 1971: Folk ideas as units of worldview. *Journal of American Folklore* 84(331).

Estonian Folklore Archives 2017: *Archives As Knowledge Hubs: Initiatives and Influence: Abstracts* Tartu.

Feilberg, Henning F 1886: *Bidrag Til En Ordbog Over Jyske Almuesmål*. Kjøbenhavn.

Franke, Simon: 1934. *Legenden Langs De Noordzee*. Zutphen.

Gunnell, Terry A. 2012: Waking the Dead: Folk Legends Concerning Magicians and Walking Corpses in Iceland. In: *News from Other Worlds: Studies in Nordic Folklore, Mythology and Culture. In Honor of John F. Lindow.* Berkeley.

Gunnell, Terry A. 2010: Sagnagrunnur: A New Database of Icelandic Folk Legends in Print. *Folklore: Electronic Journal of Folklore* 45.

Gunnell, Terry A. 2009: Legends and landscape in the Nordic countries. *Cultural and Social History* 6(3).

Gunnell, Terry A., ed. 2005: *Legends and landscape*. Reykjavik.

Hudson, Charles 1966: Folk history and ethnohistory. *Ethnohistory*.

Hult, Marte H. 2003: *Framing a National Narrative: The Legend Collections of Peter Christen Asbjørnsen*. Detroit.

Hutchins, W. John 1995: Machine translation: A brief history. In: *Concise history of the language sciences*. Oxford.

Ilyefalvi, Emese 2018: The theoretical, methodological and technical issues of digital folklore databases and computational folkloristics. *Acta Ethnographica Hungarica* 63(1).
Joosen, Vanessa 2014: *Grimms' Tales Around the Globe: The Dynamics of Their International Reception*. Detroit.
Klintberg, Bengt af 2010: *The Types of the Swedish Folk Legend*. FF Communications 300. Helsinki.
Kristensen, Evald Tang 1894: *Gamle folks fortællinger om det jyske almueliv*. Vol. 6: Vore fædres tankesæt og åndsliv. Kolding.
Kristensen, Evald Tang 1980: *Danske Sagn Som De Har Lydt in Folkemunde : Udelukkende Efter Utrykte Kilder*. København.
Krohn, Kaarle Leopold, and Julius Krohn 1926: *Die Folkloristische Arbeitsmethode*. Instituttet for Sammenlignende Kulturforskning, Serie B, Skrifter 5. Oslo.
Kverndokk, Kyrre and Hans-Jacob Ågotnes, PI 2020: *SAMLA: National Infrastructure for Cultural History and Tradition Archives*. Bergen.
Meder, Theo 2018: ISEBEL: Intelligent Search Engine for Belief Legends. *Volkskunde* 119(1).
Meder, Theo, Folgert Karsdorp, Dong Nguyen, Mariët Theune, Dolf Trieschnigg, and Iwe Everhardus Christiaan Muiser 2016: Automatic enrichment and classification of folktales in the Dutch folktale database. *Journal of American Folklore* 129(511).
Mitchell, Stephen A. 2011: *Witchcraft and Magic in the Nordic Middle Ages*. Middle Ages Series. Philadelphia.
Mitchell, Stephen A. 1997: *Blåkulla and its antecedents: transvection and conventicles in Nordic witchcraft.* Berlin.
Olrik, Axel 1910: *Dansk Folkemindesamling (Dfs): The National Collection of Folklore in Copenhagen*. FF Communications 1. Helsinki.
Saatkamp, Marielies, and Schlüter Dick 1995: *Over Heksen, Toverij En Bijgeloof in De Nederlands-Duitse Grensstreek*. Westmünsterland Bd. 4. Enschede.
Schmitt, Christoph, and Timothy R. Tangherlini 2018: Folklore Archives Online: Zur Sichtbarmachung, Auswertbarkeit und Interoperabilität einer dänischen und einer nordostdeutschen Sammlung. In: *Jahrbuch für Europäische Ethnologie Dritte Folge 13–2018*. München.
Sennrich, Rico, Barry Haddow, and Alexandra Birch 2015: Neural machine translation of rare words with subword units. arXiv:1508.07909
Sinclair, Stéfan and Geoffrey Rockwell 2016: Voyant Tools. Web. http://voyant-tools.org/.
Singh, Jaspreet, Wolfgang Nejdl, and Avishek Anand 2016: History by diversity: Helping historians search news archives. In: *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*.
Skott, Fredrik 2017: Sägenkartan. http://www.sprakochfolkminnen.se/om-oss/kartor/sagenkartan.html#/places
Stokker, Kathleen 1995: Between Sin and Salvation: The Human Condition in Legends of the Black Book Minister. *Scandinavian Studies* 67(1).
Storm, Ida, and Timothy R. Tangherlini 2018: 'En temmelig lang fodtur': hGIS, Text Mining, and Folklore Collection in 19th Century Denmark. *Human IT* 14(2).
Sturtevant, William C. 1966: "Anthropology, history, and ethnohistory." *Ethnohistory*.
Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le 2014: Sequence to sequence learning with neural networks. In: *Advances in neural information processing systems*.
Tangherlini, Timothy R. 2018: Toward a generative model of legend: Pizzas, bridges, vaccines, and witches. *Humanities* 7(1).
Tangherlini, Timothy R. 2013a: *Danish Folktales, Legends, and Other Stories*. Seattle.
Tangherlini, Timothy R. 2013b: The Folklore Macroscope: Challenges for a Computational Folkloristics. *Western Folklore* 72.
Tangherlini, Timothy R. 2010: Legendary Performances. *Ethnologia Europaea* 40(2).
Tangherlini, Timothy R. 2005: The Beggar, the Minister, the Farmer, his Wife and the Teacher: Legend and Legislative Reform in Nineteenth-Century Denmark. In: *Legends and Landscape*. Reykjavik.
Tangherlini, Timothy R. 2000: 'How do you know she's a witch?': Witches, Cunning Folk, and Competition in Denmark. *Western Folklore* 59(3/4).
Tangherlini, Timothy R 1998: 'Who ya gonna call?': Ministers and the Mediation of Ghostly Threat in

Danish Legend Tradition. *Western Folklore* 57(2/3).
Tangherlini, Timothy R. 1994: *Interpreting Legend: Danish Storytellers and Their Repertoires*. Milman Parry Studies in Oral Tradition. New York.
Tangherlini, Timothy R., and Peter M. Broadwell 2014: Sites of (re) Collection: Creating the Danish folklore nexus. *Journal of Folklore Research* 51(2).
Tangherlini, Timothy R., and Peter Leonard 2013: Trawling in the Sea of the Great Unread: Sub-corpus topic modeling and Humanities research. *Poetics* 41(6).
Taussig, Michael T. 2010: *The Devil and Commodity Fetishism in South America*. Chapel Hill.
Thompson, Stith 1966: *Motif Index of Folk-Literature*. Bloomington.
Tiedemann, Jörg 2016: OPUS--Parallel Corpora for Everyone. *Baltic Journal of Modern Computing*.
Trieschnigg, Dolf, Djoerd Hiemstra, Mariët Theune, Franciska Jong, and Theo Meder 2012: An Exploration of Language Identification Techniques in the Dutch Folktale Database. In: *Proceedings of the Workshop on Adaptation of Language Resources and Tools for Processing Cultural Heritage (LREC 2012)*.
Trojahn, Cássia, Bo Fu, Ondřej Zamazal, and Dominique Ritze 2014: State-of-the-art in multilingual and cross-lingual ontology matching. In: *Towards the Multilingual Semantic Web*. Berlin.
Uther Hans-Jörg 2004: *The Types of International Folktales*. FF Communications No. 284–286. Helsinki.
Van de Sompel, Herbert, Michael L. Nelson, Carl Lagoze, and Simeon Warner 2004: Resource harvesting within the OAI-PMH framework. *D-lib magazine* 10(12).
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I. 2017: Attention is all you need. In: *Advances in neural information processing systems*.